Normované normálne rozdelenie

Skôr ako sa budeme venovať vlastnému odhadu strednej hodnoty, zoznámime sa so základným – normovaným normálnym rozdelením N(0,1).

Funkcia hustoty f rozdelenia N(0,1) je daná predpisom:

```
f=inline('exp(-x.^2/2)/sqrt(2*pi)');
```

Pre prácu s normálnym rozdelením potrebujeme distribučnú funkciu F, teda integrál z f. Nachádza sa v matlabovskom štatistickom toolboxe pod menom *normcdf*. Pokiaľ štatistický balík nie je k dispozícii, musíme si poradiť inak. Zdanlivým problémom je nevyjadriteľnosť funkcie F za pomoci bežných „kalkulačkových“ funkcií, ale sada matlabovských funkcií je širšia a integrovanie funkcie f umožňuje – kľúčovým pomocníkom je funkcia erf, ktorú definujeme ako integrál jednej z takých „neintegrovateľných“ funkcií:

$$\mathrm{erf}(x) = \int_{[0,x]} 2/\sqrt{\pi} * \exp(-t^2)\, dt$$

Funkciu F normálneho rozdelenia by sme aj „ručne“ mali vedieť vyjadriť pomocou erf, avšak bude poučné sledovať, ako si s tým poradí Matlab. Prikážeme mu integrovať vyššie zadanú f:

```
>> s=sym('s')
>> F=int(f(s))
    F = 1125899906842624/5644425081792261*pi^(1/2)*2^(1/2)*erf(1/2*2^(1/2)*s)
```

Symbolický mód sa nezdržuje úpravami získaného výsledku, je to teda na nás:

```
>> 1125899906842624/5644425081792261*pi^(1/2)*2^(1/2)
        ans =    0.50000000000000
```

F môžeme teda predbežne (!) zapísať nasledovne:

```
>> F=inline('0.5*erf(sqrt(0.5)*x)')
```

Pri neurčitom integrovaní dostávame výsledok F+c, kde konštantu c musíme spresniť na základe iných dostupných údajov. Zatiaľ platí:

```
>> F(0)
                ans =     0
```

My však potrebujeme mať v nule hodnotu 0.5 (podľa definície), preto musíme ku F pridať 0.5. Distribučná funkcia normovaného normálneho rozdelenia potom bude

```
>> F=inline('0.5*erf(sqrt(0.5)*x)+0.5');
```

Okrem F budeme potrebovať aj jej inverznú (kvantilovú) funkciu. Tú nájdeme výpočtom na papieri, pričom využijeme matlabovskú erfinv (inverzná k erf):

```
>> Finv=inline('sqrt(2)*erfinv(2*y-1)')
```

Pozor – do Finv smieme dosadzovať iba hodnoty medzi 0 a 1 (prečo?).

Materiál na ďalšie štúdium: http://en.wikipedia.org/wiki/Normal_distribution
http://en.wikipedia.org/wiki/Error_function

Intervalový odhad strednej hodnoty

Nech X má rozdelenie N(m, 0.3^2). Realizujme náhodný výber v rozsahu n = 30:

```
>> x = rand(1,30)*25;
```

Bodový odhad strednej hodnoty m je

```
>> xm = mean(x)

xm =
   12.7693217282068
```

Hľadajme intervalový odhad strednej hodnoty so spoľahlivosťou p=0.8, teda 80%. Na to potrebujeme najprv zistiť číslo *u*, o ktorom pri normovanom N(0,1) platí:

$$quad(f, -u, u) = 0.8 ,$$

A teraz k samotnému výpočtu. Ak hľadáme také u, aby bolo quad(f, -u, u) = 0.8, potrebujeme vlastne zistiť, kedy je F(u)=0.9 (prečo je to tak?):

```
>> u=Finv(.9)
                    u =   1.28155156554460
```

Zistili sme teda hodnotu *u* a máme interval [-*u*, *u*], ktorý zodpovedá spoľahlivosti 0.8 pri normovanom normálnom rozdelení. Vrátime sa teraz k nášmu rozdeleniu N(m, 0.3^2) a zistíme, ako bude vyzerať hľadaný interval [xm-d, xm+d]. Podľa teórie (prednáška, skriptá) vieme určiť d = σ * u / sqrt(n), čo je u nás:

```
>> d = 0.3*u/sqrt(30)

        d =   0.07019347010548
```

Pri požadovanej spoľahlivosti 0.8 teda možno hľadať strednú hodnotu m na intervale

```
>> [xm-d, xm+d]

        ans =  12.69912825810132  12.83951519831229
```

Úlohy:

1. Riešte vyššie uvedený problém (tj. odhad m) pre požadovanú spoľahlivosť 0.9, 0.95 a 0.999. To isté pri východzom rozdelení N(m, 0.1^2) a N(m, 2.5^2). Ako sa mení interval odhadu v závislosti od požadovanej spoľahlivosti a od sigmy východzieho rozdelenia? Správnosť svojej odpovede testujte aj na ďalších hodnotách.
2. Riešte opačný problém – hodnota d je známa a zisťujeme spoľahlivosť *p*.
3. Zostrojte m-file, ktorý po zadaní vstupných údajov vypočíta hľadaný odhad *m*. Tj. pôjde o zhrnutie na jednom mieste celý vyššie uvedený proces hľadania.

Studentovo rozdelenie

Studentovo rozdelenie s koeficientom voľnosti n nájdeme podrobne charakterizované na stránkach:

http://en.wikipedia.org/wiki/Student_distribution

Pre lepšie porozumenie témy odporúčame tiež stránky:

http://en.wikipedia.org/wiki/Beta_function
http://en.wikipedia.org/wiki/Incomplete_beta_function#Incomplete_beta_function

Pre nás je podstatné to, že vieme distribučnú funkciu Fk (k je stupeň voľnosti) vyjadriť v tvare:

$$F(k,t) = 1 - 0.5 * I_{x(t)}(k/2 , 1/2)$$

kde $x(t) = k / (k+t^2)$.

Tzv. *regularizovanú nekompletnú beta-funkciu* $I_x(a,b)$ Matlab **pozná** pod menom betainc(x,a,b) . Pre stupeň voľnosti k teda môžeme zadať predpis distribučnej funkcie Studentovho rozdelenia jednoducho v tvare:

```
F = inline('1-0.5*betainc(k./(k+t.^2),k/2,0.5)')
```

Poznámka: Pre väčšie k Studentovo konverguje k normovanému Gaussovmu rozdeleniu.
Úloha: Nakreslite cez seba grafy N(0,1) a Studentovho rozdelenia s k=30, 40, 50. Porovnajte (aj pod lupou).

Neznámy rozptyl

Ak rozptyl σ^2 nepoznáme, nahradíme σ odhadom S=sqrt(var(x)).

```
>> S=sqrt(var(x))
                    S =    7.818833709393946
```

Ak miesto presnej sigmy používame odhad S, „zaplatíme za to“ tým, že interval [-*u*,*u*] normovanej premennej budeme hľadať (namiesto normálneho rozdelenia) zo Studentovho rozdelenia stupňa voľnosti 29. Pre spoľahlivosť 0.8 potrebujeme také *u*, v ktorom by platilo F(29,u)=0.9 (prečo 0.9?).

Najprv skusmo niekoľkými náhodnými dosadeniami zistíme, že sa nachádza niekde medzi 1 a 2. Tento odhad je potrebný na to, aby sme vedeli „poradiť“ funkcii fzero, kde má hľadať riešenie.

Matlabovská fzero hľadá nulové body (korene) funkcií. Ak ju chceme použiť na hľadanie bodu s hodnotou 0.9, musíme vlastne hľadať korene rovnice F(29,u) – 0.9 = 0. Aby sme zbytočne nekomplikovali situáciu, stupeň k=29 dosadíme „natvrdo“:

```
>> F09 = inline('1-0.5*betainc(29./(29+t.^2),29/2,0.5)-0.9')
>> format long
>> u=fzero(F09,1)
        u =   1.311433647301550
```

Hoci zobrazujeme vo formáte long, je pravdepodobné, že na posledné cifry zápisu nemusí byť spoľahnutie.

Pokračujme prechodom k nášmu konkrétnemu prípadu:

```
>> d = S*u/sqrt(30)
                    d = 1.872094086440487

>> [xm-d, xm+d]
                    10.897227641766316  14.641415814647290
```

Hľadané m je na 80% na intervale [10.897227641766316 , 14.641415814647290].

Úlohy:

1. Riešte vyššie uvedený problém (tj. odhad m) pre požadovanú spoľahlivosť 0.9, 0.95 a 0.999. Ako sa mení interval odhadu v závislosti od požadovanej spoľahlivosti?
2. Riešte opačnú úlohu – známe d a neznáme p.
3. Nájdite (a zdôvodnite) hodnotu k, pri ktorej sa (vzhľadom na presnosť Matlabu) už náhrada Studenta Gaussom nijak výraznejšie neprejaví na kvalite výsledkov.